PAC-Bayes-Chernoff bounds for unbounded losses

Ioar Casado[1], Luis A. Ortega[2], Aritz Pérez[1], Andrés R. Masegosa[3]

[1] Basque Center for Applied Mathematics (BCAM)
[2] Universidad Autónoma de Madrid (UAM)
[3] Aalborg University

Vancouver (Canada), 9-15 December 2024

Introduction

PAC-Bayes theory provides high-probability generalization bounds for aggregated predictors.

- Instead of learning a single predictor, we are interested on a probability measure $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ over the set of candidate predictors $\boldsymbol{\Theta}$.

Notation:

- Data, $D = \{\mathbf{x}_i\}_{i=1}^n$, is i.i.d. generated from an unknown distribution, ν, with support on $\mathcal{X}$
- We have a loss function $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}_+$
- *Population risk* of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as $L(\boldsymbol{\theta}) := \mathbb{E}_\nu[\ell(\boldsymbol{\theta}, \mathbf{X})]$
- *Empirical risk* of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as $\hat{L}(\boldsymbol{\theta}, D) := \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, \mathbf{x}_i)$

Standard PAC-Bayes bounds for bounded losses (McAllester, 2003):

Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D. Then,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \sqrt{\frac{KL(\rho|\pi) + \log \frac{2\sqrt{n}}{\delta}}{2n}}\,,$$

- The inequality holds simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ with probability no less than $1 - \delta$ over the choice of $D \sim \nu^n$.
- We can minimize the bound with respect to ρ to obtain novel learning algorithms.

Unbounded losses are widely used in machine learning (e.g., cross-entropy, MSE).

- PAC-Bayes bounds for unbounded losses involve extra difficulties.
- Most existing PAC-Bayes bounds for unbounded losses are derived from next result:

PAC-Bayes (oracle) bound for unbounded losses

Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D, Then, for any $\delta \in (0,1)$ and any $\lambda > 0$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \frac{1}{\lambda}\left[\frac{\mathrm{KL}(\rho|\pi) + \log\frac{f_{\pi,\nu}(\lambda)}{\delta}}{n}\right],$$

where $f_{\pi,\nu}(\lambda) := \mathbb{E}_{\pi}\,\mathbb{E}_{\nu^n}\left[e^{\lambda n\,(L(\boldsymbol{\theta})-\hat{L}(D,\boldsymbol{\theta}))}\right]$.

[Alquier, P., Ridgway, J., & Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. Journal of Machine Learning Research, 17(236), 1-41.]

Main difficulties:

- The exponential moment term $f_{\pi,\nu}(\lambda)$ has to be bounded using extra assumptions on the loss (e.g., sub-Gaussian assumption).
- The free parameter $\lambda > 0$ cannot be exactly optimized (discrete grid + union bounds).

Contributions

- A **novel PAC-Bayes oracle bound** for unbounded losses
  - Extends classic Cramér-Chernoff bounds to the PAC-Bayesian setup.
- Provides a **general framework to obtain empirical bounds** where:
  - **The free parameter λ is exactly optimized** without resorting to union-bound approaches.
  - The exponential moment term is averaged by the posterior, resulting in **more informative generalization bounds**.
  - Can be minimized to obtain **novel posteriors**.
- We illustrate the framework in several cases: generalized sub-Gaussian losses, L2 regularization, and input-gradient regularization.

Main Theorem

PAC-Bayes-Chernoff (oracle) bound

Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D,\boldsymbol{\theta})] + \inf_{\lambda \in [0,b)} \left\{ \frac{\mathrm{KL}(\rho|\pi) + \log \frac{n}{\delta}}{\lambda(n-1)} + \frac{\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]}{\lambda} \right\}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$, where $\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathbb{E}_{\nu}\left[e^{\lambda\,(L(\boldsymbol{\theta}) - \ell(\mathbf{x},\boldsymbol{\theta}))} \right]$ is the Cumulant Generating Function (CGF) of the loss.

Observations:

- Parameter-free bound without union-bounds at a $\log n$ cost.
- The risk of $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$ depends on a three-way trade-off:
  - The empirical risk $\mathbb{E}_{\rho}[\hat{L}(D,\boldsymbol{\theta})]$
  - The KL term $\mathrm{KL}(\rho|\pi)$
  - **[Novel term]** The averaged CGF term $\mathbb{E}_{\rho}[\Lambda_{\boldsymbol{\theta}}(\lambda)]$
- If the loss is the 0-1 loss, we recover Langford-Seeger's bound.
  [Seeger, Matthias. "PAC-Bayesian generalisation error bounds for Gaussian process classification." Journal of Machine Learning Research 3.Oct (2002): 233-269.]

Beyond bounded CGFs

In order to get rid of the exponential moment term, the standard practice is to make assumptions on the tails of the loss function, such as the bounded CGF assumption (which generalizes the sub-Gaussian, sub-gamma and sub-exponential cases).

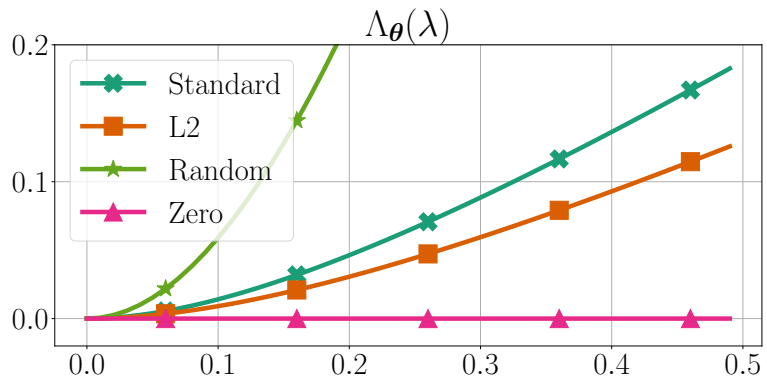
[Rodríguez-Gálvez, Borja, Ragnar Thobaben, and Mikael Skoglund. "More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity." Journal of Machine Learning Research 25.110 (2024): 1-43.]

Definition (bounded CGF)

A loss function ℓ is of bounded CGF if there is a convex and continuously differentiable function $\psi(\lambda)$ such that $\psi(0) = \psi'(0) = 0$ and for for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $\lambda \geq 0$,

$$\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathop{\mathbb{E}}_{\nu} \left[e^{\lambda \left(L(\boldsymbol{\theta}) - \ell(\mathbf{X}, \boldsymbol{\theta}) \right)} \right] \leq \psi(\lambda) \, .$$

However, uniformly bounding the CGF of every model discards information on how well each loss concentrates, which varies a lot among models, resulting in a worst-case bound.



CGFs of LeNet5 models trained with different regularization techniques.

Beyond bounded CGFs

In order to better exploit the differences between models, we introduce a generalization of the bounded CGF assumption:

Definition (Model-dependent bounded CGF)

A loss function ℓ has model-dependent bounded CGF if for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is a convex and continuously differentiable function $\psi(\boldsymbol{\theta}, \lambda)$ such that $\psi(\boldsymbol{\theta}, 0) = \psi'(\boldsymbol{\theta}, 0) = 0$ and for all $\lambda \geq 0$,

$$\Lambda_{\boldsymbol{\theta}}(\lambda) := \log \mathop{\mathbb{E}}_{\nu} \left[e^{\lambda \left(L(\boldsymbol{\theta}) - \ell(\mathbf{x}, \boldsymbol{\theta}) \right)} \right] \leq \psi(\boldsymbol{\theta}, \lambda) \,. \tag{1}$$

This results in more general bounds where the generalization of the posterior also relies on focusing on models whose loss is more concentrated:

Theorem

Let ℓ be a loss function with model-dependent bounded CGF. Let $\pi \in \mathcal{M}_1(\boldsymbol{\Theta})$ be any prior independent of D. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over draws of $D \sim \nu^n$,

$$\mathop{\mathbb{E}}_{\rho}[L(\boldsymbol{\theta})] \leq \mathop{\mathbb{E}}_{\rho}[\hat{L}(D, \boldsymbol{\theta})] + \inf_{\lambda \in [0, b)} \left\{ \frac{\mathrm{KL}(\rho | \pi) + \log \frac{n}{\delta}}{\lambda (n - 1)} + \frac{\mathbb{E}_{\rho}[\psi(\boldsymbol{\theta}, \lambda)]}{\lambda} \right\}$$

simultaneously for every $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$.

Remarkably, the bound above can be minimized with respect to $\rho \in \mathcal{M}_1(\boldsymbol{\Theta})$:

Theorem

Under the model-dependent bounded CGF assumption, the posterior distribution minimizing the PAC-Bayes-Chernoff bound is of the form

$$\rho^*(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left\{-(n-1)\lambda \hat{L}(D, \boldsymbol{\theta}) - (n-1)\psi(\boldsymbol{\theta}, \lambda)\right\}, \tag{2}$$

where $\lambda > 0$ has been fixed.

Observe that under this optimal posterior, the *maximum a posteriori* (MAP) estimate is

$$\boldsymbol{\theta}_{\text{MAP}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \left\{ \hat{L}(D, \boldsymbol{\theta}) + \frac{1}{\lambda}\psi(\boldsymbol{\theta}, \lambda) - \frac{1}{\lambda(n-1)} \ln \pi(\boldsymbol{\theta}) \right\},$$

where the extra term, $\psi(\boldsymbol{\theta}, \lambda)$, can be understood as a regularizer.

If you want to know more about how we apply these techniques to obtain novel empirical PAC-Bayes bounds for L2 and input-gradient regularization under Lipschitz and log-Sobolev assumptions, we invite you to read the full paper:

https://arxiv.org/pdf/2401.01148